J M S K
Jurnal Matematika, Statistika & Komputasi
http://journal.unhas.ac.id/index.php/jmsk
e-ISSN: 2614-8811
Vol .15, No. 2, 44-54, January 2019

Penerapan *Sparse Principal Component Analysis* dalam Menghasilkan Matriks *Loading* yang *Sparse*

Retno Mayapada*, Georgina M. Tinungki[1], Nurtiti Sunusi[2]

Abstract

Sparse Principal Component Analysis (*Sparse* PCA) is one of the development of PCA. *Sparse* PCA modifies new variables as a linier combination of *p* old variables (original variable) which is yielded by PCA method. Modifying new variables is conducted by producing a *loading* yang *sparse* matrix, such that old variable which is not effective (value of *loading* is zero) able be exit from PCA. In this study, *Sparse* PCA method was applied on data of Indonesia Poverty population in 2015, that contains 13 variables and 34 observation with variable reduction such that yields 4 (four) new variables, which can explain 80.1% of total variance data. This study show, the *loading* matrix that has been yielded by using *Sparse* PCA method to become *sparse* with there exist 11 elements (*loading* value) zero entry of matrix, such that the model that has been produced to be simpler and easy to be interpreted.

Keywords: Principal Component Analysis, Sparse Principal Component Analysis, reduksi dimensi, matriks loading yang sparse

Abstrak

Sparse Principal Component Analysis (*Sparse* PCA) merupakan salah satu pengembangan dari metode PCA. *Sparse* PCA memodifikasi variabel-variabel baru yang merupakan kombinasi linear dari *p* variabel lama (variabel asli) yang dihasilkan oleh metode PCA. Pemodifikasian variabel baru ini dilakukan dengan dengan menghasilkan matriks *loading* yang *sparse* sehingga variabel lama yang tidak efektif (memiliki nilai *loading* sama dengan nol) dapat dikeluarkan dari model PCA. Pada penelitian ini, metode *Sparse* PCA diterapkan pada data Indikator Kemiskinan Penduduk Indonesia Tahun 2015 yang memuat 13 variabel dan 34 observasi dengan reduksi variabel menghasilkan 4 (empat) variabel baru yang telah mampu menjelaskan 80,1% dari total variansi data. Hasil penelitian menunjukkan, matriks *loading* yang dihasilkan menggunakan metode *Sparse* PCA menjadi *sparse* dengan terdapat 11 elemen (nilai *loading*) matriks bernilai nol sehingga model yang dihasilkan menjadi lebih sederhana dan mudah untuk diinterpretasikan.

Kata Kunci: *Principal Component Analysis, Sparse Principal Component Analysis*, reduksi dimensi, matriks *loading* yang *sparse*

1. Pendahuluan

Principal Component Analysis (PCA) atau Analisis Komponen Utama (AKU) pertama kali diperkenalkan oleh Karl Pearson pada tahun 1901. PCA digunakan untuk menghitung kombinasi linier dan variabel baru yang menggambarkan keragaman data asli sebanyak mungkin, dengan dimensi matriks data asli dapat disederhanakan tanpa harus kehilangan informasi penting (Setyaningsih *et al*., 2010). Hingga saat ini, PCA telah banyak digunakan dalam berbagai bidang penelitian.

Akan tetapi, terdapat kekurangan pada metode PCA yaitu setiap *Principal Component* (PC) atau Komponen Utama (KU) merupakan kombinasi linier dari semua *p* variabel. Artinya, setiap KU merupakan kombinasi linier dari semua variabel dengan beban (nilai *loading)*

*Program Studi Statistika, Departemen Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin
[1]ina_matematika@yahoo.co.id, ntitisanusi@gmail.com[2]

diberikan ke setiap variabel. Nilai *loading* yang dihasilkan ini biasanya tidak nol. Hal ini mengakibatkan hasil KU yang diperoleh akan sulit untuk diinterpretasikan.

Salah satu perkembangan metode PCA adalah *Sparse Principal Component Analysis* (*Sparse* PCA) yang dapat digunakan untuk mengatasi masalah ini [1]. *Sparse* PCA diperkenalkan oleh Zou *et al*. pada tahun 2004. *Sparse* PCA menggabungkan kekuatan PCA klasik, reduksi data, dan pemodelan *sparseness*, yang menghasilkan matriks loading yang *sparse* sehingga dapat mengeluarkan variabel yang tidak efektif dari model PCA yang memiliki nilai *loading* (bobot) sama dengan nol [1]. Oleh karena itu, *Sparse* PCA memiliki kelebihan dalam membuat interpretasi KU menjadi lebih mudah.

2. Landasan Teori

2.1 *Principal Component Analysis*

Secara aljabar, *Principal Component Analysis* **(**PCA) adalah kombinasi linear khusus dari p variabel acak $X_1, X_2, \dots, X_P$. Secara geometri, kombinasi linear ini menggambarkan pemilihan dari sistem koordinat yang diperoleh dengan merotasikan sistem awal dengan $X_1, X_2, \dots, X_P$ sebagai sumbu koordinat. Sumbu baru merupakan arah dengan variabilitas maksimum dan memberikan struktur kovariansi yang lebih sederhana. Prosedur PCA pada dasarnya adalah bertujuan untuk menyederhanakan variabel yang diamati dengan cara menyusutkan (mereduksi) dimensinya. Hal ini dilakukan dengan cara menghilangkan korelasi diantara variabel prediktor melalui transformasi variabel prediktor asal ke variabel baru yang tidak berkorelasi sama sekali atau yang biasa disebut dengan KU.

Setelah beberapa komponen hasil PCA yang bebas multikolinearitas diperoleh, maka komponen-komponen tersebut menjadi variabel prediktor baru yang dapat diregresikan atau dianalisa pengaruhnya terhadap variabel respon (Y) dengan menggunakan analisis regresi. Reduksi data pengamatan ke dalam beberapa set data menggunakan PCA dapat dilakukan sedemikian sehingga informasi dari semua data dapat diserap seoptimal mungkin. Oleh karena itu, PCA dapat dipandang sebagai transformasi dari $X_1, X_2, \dots, X_P$ [2]. KU yang terbentuk dapat dituliskan sebagai kombinasi linier dari variabel-variabel asalnya dan vektor eigen $\boldsymbol{e_1, e_2, \dots, e_P}$ dari matriks kovarians yang bersesuaian dengan nilai-nilai eigen seperti pada persamaan (1) berikut.

$$\text{KU}_j = e_{1j}X_1 + e_{2j}X_2 + \cdots + e_{Pj}X_p \, , j = 1,2, \dots p \qquad (1)$$

Penentuan jumlah KU yang terpilih dapat didasarkan pada 3 kriteria yaitu dengan melihat titik pada *scree plot* ketika kurva mulai landai, nilai eigen > 1, dan keragaman kumulatif yang dapat dijelaskan oleh KU minimal 80%.

2.2 Nilai *Loading*

Nilai *loading* adalah koefisien dari transformasi KU yang memberikan hasil yang tepat mengenai pengaruh variabel-variabel asli dari KU dan merupakan dasar yang bermanfaat untuk interpretasi. Nilai koefisien yang besar menerangkan *loading* yang tinggi dan ketika nilai koefisien mendekati nol, artinya KU tersebut memiliki *loading* yang rendah [3]. Variabel yang memiliki nilai *loading* tepat nol dapat dikeluarkan dari fungsi KU.

Matriks *loading* yang *sparse* adalah matriks loading yang elemen-elemen (nilai *loading*)nya banyak bernilai nol. Nilai *loading* memberikan indikasi variabel asli (lama) mana yang sangat penting atau mempengaruhi pembentukan KU sebagai variabel baru. Semakin tinggi nilai *loading* dari suatu variabel lama maka semakin besar pula pengaruhnya terhadap pembentukan variabel baru [4]. Sebaliknya, semakin rendah nilai *loading* dari suatu variabel

lama maka semakin kecil pula pengaruhnya terhadap pembentukan variabel baru. Salah satu metode estimasi nilai *loading* yaitu metode PCA.

Nilai *loading* juga merupakan nilai vektor eigen dari penduga matriks kovarians dari **X** [5]. Sehingga nilai *loading* yang diperoleh hasilnya sama dengan nilai vektor eigen. Nilai *loading* ini digunakan sebagai koefisien dari fungsi KUnya [6]. Jenis pengaruh suatu variabel berdasarkan nilai *loading*nya adalah sebagai berikut [7].

- 0 – 0.3 : Tidak berpengaruh
- >0.3 – 0.4 : Berpengaruh
- >0.4 – 0.5 : Berpengaruh, dianggap penting
- >0.5 – 1 : Berpengaruh signifikan

2.3 *Sparse Principal Component Analysis*

Salah satu bentuk pengembangan terbaru dari PCA adalah *Sparse* PCA. *Sparse* PCA menggabungkan kelebihan PCA klasik, reduksi data, dengan pemodelan *sparseness*, yang menghilangkan variabel yang tidak efektif dari model PCA dengan mengecilkan nilai *loading* dari variabel-variabel ini menjadi nol. Oleh karena itu, *Sparse* PCA memiliki kelebihan dalam membuat interpretasi KU menjadi lebih mudah.

Zou *et al*. (2004) memperkenalkan *Sparse* PCA menggunakan metode *Elastic Net* untuk menghasilkan KU yang dimodifikasi dari nilai-nilai *loading* yang *sparse*. PCA dapat diformulasikan sebagai masalah optimasi pada regresi, sehingga nilai-nilai *loading* dapat diperoleh dengan menerapkan batasan *Elastic Net* pada koefisien regresi β [8]. *Elastic Net* merupakan suatu metode seleksi variabel dengan menggabungkan regresi ridge dan LASSO. Dengan kata lain, *Elastic Net* menggabungkan batasan L_1*-norm* dan L_2*-norm* kuadrat pada β.

Penggabungan dua batasan tersebut diharapkan dapat menyeimbangkan kelemahan dari masing-masing metode (*ridge* dan LASSO) dengan batasan L_1*-norm* menghasilkan model yang lebih sederhana karena terjadi penyusutan beberapa β yang tepat nol, sedangkan batasan L_2*-norm* kuadrat menghasilkan model yang tidak menyeleksi variabel namun meningkatkan efek pengelompokan dan penyusutan β [9]. Metode *Sparse* PCA dengan menempatkan batasan L_1*-norm* dan L_2*-norm* kuadrat ini dapat menghasilkan nilai-nilai *loading* yang *sparse* dan persentase varians yang lebih tinggi daripada metode *Sparse* PCA yang hanya menempatkan batasan L_1*-norm.*

2.3.1 Pendekatan *Sparse*

Perhatikan bahwa setiap KU merupakan kombinasi linier dari p variabel, sehingga pembebanannya (nilai-nilai *loading*) dapat diperoleh dengan cara meregresikan KU pada p variabel.

Teorema 1 [12]

Untuk setiap i, $\boldsymbol{W_i}$ adalah KU ke-i. Misalkan $\mathbf{X}_{n\times p}$ $(n > p)$ adalah matriks yang memiliki rank penuh (*full-rank*), ℓ suatu bilangan non-negatif dan penduga *ridge* $\hat{\beta}_{ridge}$ diberikan oleh persamaan (2) berikut,

$$\widehat{\boldsymbol{\beta}}_{ridge} = \underset{\beta}{\arg\min} \{\|\boldsymbol{W_i} - \mathbf{X}\boldsymbol{\beta}\|^2 + \ell\|\boldsymbol{\beta}\|_2^2\} \qquad (2)$$

$\widehat{\boldsymbol{v}}$ diperoleh dengan menormalisasikan $\widehat{\boldsymbol{\beta}}_{ridge}$, $\widehat{\boldsymbol{v}} = \frac{\widehat{\boldsymbol{\beta}}_{ridge}}{\|\widehat{\boldsymbol{\beta}}_{ridge}\|}$, dengan $\widehat{\boldsymbol{v}} = \mathbf{v_j}$

Teorema 1 menunjukkan hubungan antara PCA dan metode regresi. Peregresian KU pada variabel-variabel dibahas oleh Cadima dan Jolliffe pada tahun 1995. Cadima dan Jolliffe (1995) berfokus pada pendekatan KU dengan subset dari variabel k [10]. Zou *et al*., (2004)

melakukan pendekatan KU menggunakan regresi *ridge*. Setelah normalisasi, koefisien saling bebas terhadap ℓ. Oleh karena itu penalti regresi *ridge* dalam bentuk L_2*-norm* kuadrat tidak digunakan untuk melakukan penalisasi koefisien regresi, tetapi untuk memastikan rekonstruksi KU.

Selanjutnya, tambahkan L_1*-norm* ke persamaan (2) sehingga diperoleh persamaan (3) berikut [11].

$$\widehat{\boldsymbol{\beta}} = \underset{\beta}{\arg\min}\{\|\boldsymbol{W_i} - \mathbf{X}\boldsymbol{\beta}\|^2 + \ell\|\boldsymbol{\beta}\|_2^2 + \ell_1\|\boldsymbol{\beta}\|_1\} \quad (3)$$

dan $\hat{\mathbf{v}}_\mathbf{i} = \frac{\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|}$, merupakan penduga dari $\mathbf{v_i}$ dan $\mathbf{X}\hat{\mathbf{v}}_\mathbf{i}$ adalah penduga KU ke-i. Persamaan (2.4) disebut sebagai persamaan *naive elastic net* [12] atau dengan kata lain, persamaan (2.4) adalah regularisasi *naive elastic net* yang merupakan kombinasi dari L_1*-norm* dan L_2*-norm* kuadrat.

Selanjutnya, Zou *et al*. (2006) menyusun suatu algoritma untuk meminimukan nilai *loading* yang dihasilkan metode *Sparse* PCA berdasarkan persamaan (3). Misal $\widehat{\mathbf{B}} = [\widehat{\boldsymbol{\beta}}_\mathbf{1}, \widehat{\boldsymbol{\beta}}_\mathbf{2}, \ldots \widehat{\boldsymbol{\beta}}_\boldsymbol{k}]$ merupakan matriks berukuran $p \times k$, p jumlah variabel prediktor dan k jumlah KU yang terpilih, dan $\widehat{\boldsymbol{\beta}}_\boldsymbol{j}$ suatu vektor penduga *naive elastic net*, subtitusi $\boldsymbol{W_j} = \mathbf{X}\boldsymbol{\alpha_j}$ untuk setiap $j = 1, \ldots k$, ke persamaan (3) sehingga diperoleh persamaan (4) [8].

$$\widehat{\boldsymbol{\beta}}_\boldsymbol{j} = \underset{\beta_j}{\arg\min}\left\{\left\|\mathbf{X}\boldsymbol{\alpha_j} - \mathbf{X}\boldsymbol{\beta_j}\right\|^2 + \ell\left\|\boldsymbol{\beta_j}\right\|_2^2 + \ell_{1,j}\left\|\boldsymbol{\beta_j}\right\|_1\right\}$$

$$\widehat{\boldsymbol{\beta}}_\boldsymbol{j} = \underset{\beta_j}{\arg\min}\left\{\left(\boldsymbol{\alpha_j} - \boldsymbol{\beta_j}\right)^{\mathrm{T}}(\|\mathbf{X}\|\|\mathbf{X}\|)(\boldsymbol{\alpha_j} - \boldsymbol{\beta_j}) + \ell\left\|\boldsymbol{\beta_j}\right\|_2^2 + \ell_{1,j}\left\|\boldsymbol{\beta_j}\right\|_1\right\}$$

dengan mensubtitusikan $\|\mathbf{X}\| = \sqrt{\mathbf{X^T X}}$, maka diperoleh persamaan (4).

$$\widehat{\boldsymbol{\beta}}_\boldsymbol{j} = \underset{\beta_j}{\arg\min}\left\{\left(\boldsymbol{\alpha_j} - \boldsymbol{\beta_j}\right)^{\mathrm{T}}\mathbf{X^T X}(\boldsymbol{\alpha_j} - \boldsymbol{\beta_j}) + \ell\left\|\boldsymbol{\beta_j}\right\|_2^2 + \ell_{1,j}\left\|\boldsymbol{\beta_j}\right\|_1\right\} \quad (4)$$

dengan $\left\|\boldsymbol{\beta_j}\right\|_2^2 = \sum_{i=1}^{p}\beta_{ij}^2$ dan $\left\|\boldsymbol{\beta_j}\right\|_1 = \sum_{i=1}^{p}|\beta_{ij}|$.

Catatan:
ℓ digunakan untuk semua k KU, sedangkan nilai $\ell_{1,j}$ dapat berbeda-beda untuk setiap $j = 1, \ldots k$ KU [8].

2.3.2 Pemilihan Parameter Tuning

Pemilihan $\ell_{1,j}$ pada persamaan (4) dapat dilakukan dengan menggunakan *Cross Validation* (CV). Nilai $\ell_{1,j}$ dipilih dengan memperhatikan nilai CV terkecil [9]. CV yang sebaiknya digunakan adalah 5-*fold* atau 10-*fold* karena memberikan dugaan sisaan prediksi yang mempunyai bias tinggi namun memberikan MSE kecil dan juga variansi yang lebih kecil [8]. Adapun untuk pemilihan ℓ, ditentukan oleh peneliti (dapat mempertimbangkan prinsip *parsimony*). ℓ yang dipilih adalah ℓ yang menghasilkan model matriks *loading* yang *sparse* dan proporsi keragaman kumulatif yang dihasilkan ≥ 80%.

2.3.4 Algoritma *General* SPCA

Algoritma untuk *Sparse* PCA pada data dengan $n > p$ yang disebut sebagai algoritma *General* SPCA untuk mereduksi dimensi data menggunakan penduga *Naive Elastic Net* yang akan diterapkan pada penelitian ini sebagai berikut [11].

1. Misalkan **A** berisikan $\mathbf{V}[, 1{:}\, k]$ merupakan nilai-nilai *loading* dari metode PCA
2. Diberikan $\mathbf{A} = [\boldsymbol{\alpha_1}, \ldots, \boldsymbol{\alpha_k}]$, hitung persamaan penduga *naive elastic net* untuk $j = 1,2,\ldots,k$ berikut.

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}} = \underset{\beta_j}{\arg\min} \left\{ \left(\boldsymbol{\alpha}_{\boldsymbol{j}} - \boldsymbol{\beta}_{\boldsymbol{j}}\right)^{\mathrm{T}} \mathbf{X}^{\mathbf{T}}\mathbf{X} \left(\boldsymbol{\alpha}_{\boldsymbol{j}} - \boldsymbol{\beta}_{\boldsymbol{j}}\right) + \ell \left\| \boldsymbol{\beta}_{\boldsymbol{j}} \right\|_2^2 + \ell_{1,j} \left\| \boldsymbol{\beta}_{\boldsymbol{j}} \right\|_1 \right\}$$

3. Untuk setiap $\mathbf{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k]$ yang diperoleh dari langkah 2, hitung SVD dari $\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathbf{T}}$, lalu perbaharui $\mathbf{A} = \mathbf{U}\mathbf{V}^{\mathbf{T}}$
4. Ulangi langkah 2-3 sampai β konvergen
5. Lakukan normalisasi: $\hat{\mathbf{v}}_{\mathbf{j}} = \frac{\boldsymbol{\beta}_j}{\|\boldsymbol{\beta}_j\|}$; $j = 1, \ldots, k$

3. Metode *Sparse Principal Component Analysis*

Metode *Sparse* PCA pada *paper* ini akan digunakan untuk memodifikasi matriks *loading* menjadi *sparse*. Sebelum menerapkan metode Sparse PCA, terlebih dahulu data akan dihitung matriks korelasinya untuk melihat apakan data terjadi multikolinieritas atau tidak. PCA akan berguna jika data yang diolah merupakan data yang terjadi multikolinieritas. Selanjut, data distandarisasi untuk menghilangkan satuan sehingga variabel-variabel pada data layak untuk dibandingkan. Setelah data distandarisasi, data kemudian diolah menggunakan metode PCA berdasarkan matriks kovarians untuk menentukan jumlah KU yang terpilih yaitu sebanyak *k* dan menghitung matriks *loading*nya.

Selanjutnya, matriks *loading* yang dihasilkan menggunakan metode PCA akan dimodifikasi menjadi matriks *loading* yang sparse menggunakan metode Sparse PCA. Pemodifikasian ini dilakukan dengan menghitung penduga *naive elastic net* $\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}$ menggunakan algoritma *General* SPCA untuk masing-masing KU. Sehingga, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}$ merupakan vektor berukuran $p \times 1$ *untuk* $j = 1,2,..,k$ yang berisi nilai-nilai loading yang *sparse*.

4. Hasil dan Pembahasan

4.1 Sumber Data

Data yang digunakan pada penelitian ini adalah data sekunder yaitu data Indikator Kemiskinan Penduduk Indonesia Tahun 2015 yang diperoleh dari Badan Pusat Statistika pada situs http://www.bps.go.id yang terdiri atas 13 variabel dan 34 observasi [13]. Variabel yang digunakan dalam penelitian ini adalah sebagai berikut.

X_1 : Tingkat Partisipasi Angkatan Kerja (TPAK) (%)
X_2 : Kepadatan Penduduk (Jiwa/km^2)
X_3 : Indeks Pembangunan Manusia (IPM)
X_4 : Jumlah Penduduk Berusia 15-44 Tahun yang Buta Huruf (%)
X_5 : Jumlah Rumah Tangga yang Memiliki Sumber Penerangan Listrik (%)
X_6 : Jumlah Rumah Tangga yang Memiliki Luas Hunian per kapita $\geq$ 7.2 m^2 (%)
X_7 : Angka Harapan Hidup (Tahun)
X_8 : Jumlah Penduduk yang Mempunyai Keluhan Kesehatan Selama Sebulan Terakhir (%)
X_9 : Jumlah Balita yang Pernah Mendapat Imunisasi Campak (%)
X_{10}: Jumlah Rumah Tangga yang Memiliki Fasilitas Tempat Buang Air Besar Sendiri (%)
X_{11}: Tingat Pengangguran Terbuka (%)
X_{12}: Pengeluaran Bahan Makanan per Kapita Sebulan (%)
X_{13}: Pengeluaran Bukan Bahan Makanan per Kapita Sebulan (%)

4.2 Matriks Korelasi

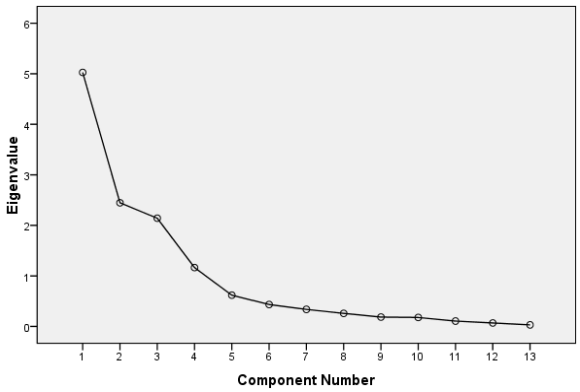
Perhitungan matriks korelasi dilakukan dengan tujuan untuk mengetahui korelasi antar variabel karena pada PCA mensyaratkan bahwa diantara variabel-variabelnya terdapat korelasi. Matriks korelasi yang diperoleh menunjukkan bahwa terdapat beberapa pasang variabel dengan nilai korelasi sangat kuat yaitu variabel X_2 dengan variabel X_{18} $(\rho_{2,18} = 0{,}771)$, X_3 dengan variabel X_7 $(\rho_{3,7} = 0{,}783)$, dan X_{12} dengan variabel X_{13} $(\rho_{12,13} = 0{,}757)$. Hal ini menunjukkan bahwa PCA dapat digunakan untuk mereduksi variabel-variabel yang saling berkorelasi.

4.3 Penentuan Jumlah Komponen Utama yang Terbentuk menggunakan *Principal Component Analysis*

Untuk menentukan jumlah KU yang terbentuk, terdapat 3 kriteria yang bisa digunakan yaitu dengan melihat scree plot, nilai eigen, dan keragaman kumulatif yang dapat dijelaskan oleh KU.

1. Scree Plot

Scree plot adalah plot antara nilai eigen dengan banyaknya KU yang terbentuk. Jumlah KU yang terbentuk dapat diketahui dengan memperhatikan patahan siku dari *scree plot. Scree Plot* data dengan metode PCA dapat dilihat pada Gambar 4.1.



Gambar 1. *Scree Plot* data dengan metode PCA

Pada Gambar 1, terlihat bahwa kurva mulai meluruh setelah KU ke-4. Sehingga dapat disimpulkan bahwa jumlah KU yang terpilih adalah 4 KU.

2. Nilai eigen

KU yang terbentuk dapat ditentukan dengan memilih KU yang memiliki nilai eigen lebih besar dari 1. Nilai eigen dengan metode PCA disajikan dalam Tabel 1 dengan nilai eigen tiap KU telah terurut. Artinya KU pertama memiliki nilai eigen paling besar dan KU ke-tiga belas memiliki nilai eigen paling kecil.

Tabel 1. Nilai Eigen dengan Metode PCA

KU ke-	Nilai Eigen	KU ke-	Nilai Eigen
1	**5.025562**	8	0.259007
2	**2.44679**	9	0.18576
3	**2.141998**	10	0.177708
4	**1.164892**	11	0.106607

5	0.619113	12	0.068969
6	0.434757	13	0.031239
7	0.337597		

Sumber: Data Diolah, 2018

Tabel 1 menunjukkan bahwa terdapat 4 KU yang memiliki nilai eigen lebih besar dari 1. KU_1 memiliki nilai eigen sebesar 5.025562. Selanjutnya, KU_2 memiliki nilai eigen sebesar 2.44679, KU_3 memiliki nilai eigen sebesar 2.141998 dan KU_4 memiliki nilai eigen sebesar 1.164892. Adapun KU_5 hingga KU_{13} memiliki nilai eigen kurang dari 1. Sehingga dapat disimpulkan bahwa jumlah KU yang terbentuk menurut kriteria nilai eigen adalah 4 KU.

3. Persentase Proporsi Keragaman Kumulatif

KU yang terbentuk dapat ditentukan dengan memilih KU yang dapat menjelaskan proporsi keragaman data secara kumulatif minimal 80%. Tabel 2 menunjukkan bahwa secara kumulatif, keempat KU pertama telah mampu menjelaskan 82.92% dari total variansi data. Sehingga, jumlah KU yang terbentuk menurut criteria ini adalah 4 KU.

Tabel 2. Proporsi Keragaman dan Keragaman Kumulatif dengan Metode PCA

Komponen Utama ke-	Proporsi keragaman	Proporsi keragaman kumulatif
1	0.3866	**0.3866**
2	0.1882	**0.5748**
3	0.1648	**0.7396**
4	0.0896	**0.8292**
5	0.0476	0.8768
6	0.0334	0.9102
7	0.0260	0.9362
8	0.0199	0.9561
9	0.0143	0.9704
10	0.0137	0.9841
11	0.0080	0.9923
12	0.0053	0.9976
13	0.0024	1.0000

Sumber: Data Diolah, 2018

Jumlah KU yang dipilih adalah sebanyak empat (4) berdasarkan ketiga kriteria pemilihan jumlah KU yaitu *scree plot*, nilai eigen dan persentase keragaman kumulatif.

4.4 Matriks *Loading* menggunakan Metode *Principal Component Analysis*

Jumlah KU yang diperoleh berdasarkan *scree plot*, nilai eigen, dan persentase keragaman kumulatif adalah empat KU yang merupakan variabel baru. Nilai *Loading* pada Tabel 3 menjelaskan hubungan (korelasi) antara variabel lama dengan variabel baru yang dibentuk dengan metode PCA. Berdasarkan Tabel 3, nilai *loading* beberapa variabel tidak memiliki perbedaan yang signifikan sehingga sulit untuk menentukan variabel mana yang paling banyak menjelaskan tiap KU. Oleh karena itu, tahapan analisis yang digunakan untuk mengatasi masalah tersebut adalah dengan menerapkan metode *sparse* PCA yang memodifikasi matriks *loading* dari Tabel 3.

Tabel 3. Matriks *Loading* Menggunakan Metode PCA

Variabel	KU_1	KU_2	KU_3	KU_4
X_1	-0.251	0.013	-0.437	0.286

X_2	0.188	0.322	-0.350	-0.427
X_3	0.409	0.016	-0.127	-0.037
X_4	-0.304	0.048	-0.300	0.205
X_5	0.373	-0.225	0.119	-0.149
X_6	-0.243	0.409	-0.217	-0.288
X_7	0.350	-0.079	-0.120	0.232
X_8	0.107	-0.368	-0.335	-0.387
X_9	0.231	-0.300	-0.308	0.010
X_{10}	0.323	0.082	-0.014	0.505
X_{11}	0.143	0.299	0.462	-0.234
X_{12}	0.200	0.472	-0.021	0.270
X_{13}	0.303	0.013	-0.437	0.286

Sumber: Data Diolah, 2018

4.5 Pemilihan Parameter Tuning ℓ dan $\ell_{1,j}$

Pemilihan model terbaik penduga *Elastic Net* pada metode *Sparse* PCA dapat menggunakan *Cross Validation* (CV) dengan memilih $\ell_{1,j}$ yang memiliki nilai CV terkecil. Pada penelitian ini, diuji beberapa nilai ℓ untuk mendapatkan model yang sederhana namun tetap mampu menjelaskan keragaman data asli minimal 80%. Nilai ℓ yang digunakan pada penelitian ini adalah $\ell = 4$. Sedangkan untuk ℓ_1 pada model KU_1 yang terpilih 0.363636, ℓ_1 pada model KU_2 yang terpilih adalah 0.525253, ℓ_1 pada model KU_3 yang terpilih adalah 0.8485, dan ℓ_1 pada model KU_4 yang terpilih adalah 1.

4.6 Matriks *Loading* menggunakan Algoritma *Sparse Principal Component Analysis*

Sparse PCA mengeluarkan variabel yang tidak efektif dari model PCA dengan mengecilkan nilai *loading* menjadi nol. Tabel 4 merupakan hasil dari matriks *loading* menggunakan algoritma *Sparse* PCA. Berdasarkankan Tabel 4, terdapat 11 nilai *loading* yang bernilai nol, sehingga terlihat bahwa matriks *loading* yang dihasilkan oleh metode *Sparse* PCA menjadi lebih sederhana. Peneliti lebih memilih model yang lebih sederhana karena dapat memberikan kemudahan dalam melihat hubungan antara KU yang terbentuk dengan variabel lama (variabel prediktor).

Berdasarkan Tabel 4 akan dibentuk 4 KU sebagai kombinasi linier variabel-variabel asalnya yaitu sebagai berikut.

$$K_i = \boldsymbol{v'_i X} = v_{1i}X_1 + v_{2i}X_2 + \cdots + v_{pi}X_p \qquad i = 1,2,3$$

$$\mathrm{KU}_{\mathbf{1}} = \boldsymbol{v'_1 Z} = 0.111Z_1 - 0.238Z_2 - 0.423Z_3 + 0.215Z_4 - 0.372\,Z_5 + 0.220Z_6 - 0.362Z_7 - 0.302Z_8 - 0.370Z_9 - 0.249Z_{10} + 0.034Z_{11} - 0.063Z_{12} - 0.304Z_{13}$$

$$\mathrm{KU}_{\mathbf{2}} = \boldsymbol{v'_2 Z} = -0.103Z_1 - 0.470Z_2 - 0.067Z_3 - 0.091Z_4 + 0.195\,Z_5 - 0.468Z_6 + 0.184Z_8 + 0.147Z_9 - 0.055Z_{10} - 0.134Z_{11} - 0.454Z_{12} + 0.464Z_{13}$$

$$\mathrm{KU}_3 = \boldsymbol{v'_3 Z} = 0.555Z_1 + 0.424Z_4 - 0.207Z_5 + 0.081Z_7 + 0.153Z_8 + 0.256Z_9 + 0.080Z_{10} - 0.606Z_{11}$$

$$\mathrm{KU}_{\mathbf{4}} = \boldsymbol{v'_4 Z} = 0.376Z_2 + 0.278Z_6 - 0.211Z_7 + 0.525Z_8 + 0.080Z_9 - 0.555Z_{10} - 0.009Z_{11} - 0.383Z_{12}$$

Tabel 4 Matriks *Loading* menggunakan Metode *Sparse* PCA

Variabel	KU_1	KU_2	KU_3	KU_4
X_1	0.111	-0.103	0.555	0

X_2	-0.238	-0.470	0	0.376
X_3	-0.423	-0.067	0	0
X_4	0.215	-0.091	0.424	0
X_5	-0.372	0.195	-0.207	0
X_6	0.220	-0.468	0	0.278
X_7	-0.362	0	0.081	-0.211
X_8	-0.302	0.184	0.153	0.525
X_9	-0.370	0.147	0.256	0.080
X_{10}	-0.249	-0.055	0.080	-0.555
X_{11}	0.034	-0.134	-0.606	-0.009
X_{12}	-0.063	-0.454	0	-0.383
X_{13}	-0.304	-0.464	0	0

Sumber: Data diolah, 2018

Tabel 5 Pengelompokkan Variabel Lama, dan Variansi yang dijelaskan Tiap KU Menggunakan *Sparse* PCA

Komponen Utama	**Variabel Lama**	**Variansi yang dijelaskan**
KU_1	IPM (X_3) Listrik (X_5) AH (X_7) Imunisasi (X_9)	35.7%
KU_2	Kepadatan (X_2) Luas Hunian (X_6) Pengeluaran Makanan (X_{12}) Pengeluaran Non Makanan (X_{13})	18.5%
KU_3	TPAK (X_1) Buta Huruf (X_4) TPT (X_{11})	16.0%
KU_4	Keluhan Kesehatan (X_8) TBAB Sendiri (X_{10})	9.9%

Sumber: Data diolah, 2018

Pada Tabel 5 diketahui bahwa KU_1 memiliki nilai persentase varian sebesar 35.7%. Variabel yang paling banyak membentuk KU_1 yaitu variabel IPM, Listrik, Angka Harapan Hidup, dan Imunisasi KU_1 memiliki variansi paling besar diantara KU yang terbentuk lainnya, sehingga dapat dikatakan bahwa jumlahan dari variabel IPM, Listrik, Angka Harapan Hidup, dan Imunisasi menghasilkan variansi yang besar. KU_2 dapat menjelaskan 18.5% dari total varians. Variabel yang paling banyak membentuk KU_2 adalah Kepadatan, Luas Hunian, Pengeluran Makanan, dan Pengleuaran Non Makanan. Selanjutnya KU_3 mampu menjelaskan variansi sebesar 16,0% dan variabel yang paing banyak membentuk KU_3 yaitu TPAK, Buta Huruf, dan TPT. Selanjutnya KU_4 mampu menjelaskan variansi sebesar 9,9% dan variabel yang paling banyak membentuk KU_4 yaitu Keluhan Kesehatan dan TBAB Sendiri.

4.8 Penentuan Model Terbaik

Hasil analisis dari metode PCA dan *Sparse* PCA dapat dilanjutkan ke tahapan regresi untuk melihat metode mana yang dapat menghasilkan model yang lebih baik. Pada penelitian ini, skor komponen utama yang dibentuk dari matriks *loading* diregresikan dengan variabel respon. Variabel respon yang digunakan adalah variabel Garis Kemiskinan. Adapun indikator yang digunakan untuk menentukan model terbaik adalah nilai AIC dan Koefisien Determinasi (R^2). Nilai R^2 dan nilai AIC dari Model regresi menggunakan Metode PCA dan *Sparse* PCA dapat dilihat pada Tabel 7.

Tabel 7 Nilai R^2 dan nilai AIC dari Model Regresi menggunakan Metode PCA dan *Sparse* PCA

No	Jenis Metode	R^2	AIC
1	PCA	55.83%	884.797
2	*Sparse* PCA	56.25%	884.475

Sumber: Data diolah, 2018

Model terbaik berdasarkan indikator R^2 dan AIC adalah model yang memiliki nilai AIC terkecil dan nilai R^2 terbesar. Tabel 4.7 menunjukkan model yang diperoleh dari matriks *loading* yang sparse menggunakan metode *Sparse* PCA lebih baik daripada model regresi yang dihasilkan dari skor KU berdasarkan matriks loading tanpa modifikasi dari metode PCA karena mampu menghasilkan model regresi dengan nilai AIC lebih kecil dan nilai R^2 yang lebih besar.

5. Kesimpulan

Berdasarkan pembahasan, diperoleh kesimpulan bahwa matriks *loading* dari hasil reduksi variabel pada data Indikator Kemiskinan Penduduk Indonesia Tahun 2015 yang berjumlah 13 (tiga belas) variabel menjadi 4 (empat) KU lebih sederhana dan *sparse* setelah dimodifikasi menggunakan metode *Sparse* PCA. Model KU yang telah dimodifikasi dari matriks *loading* menggunakan metode *Sparse* PCA menjadi lebih sederhana (*sparse*) dan mudah untuk diinterpretasikan dengan terdapat 11 nilai *loading* bernilai nol (41 nilai *loading* lainnya tidak bernilai nol). Secara kumulatif, ke-4 (empat) KU yang dibentuk dari matriks loading yang *sparse* telah mampu menjelaskan 80.1% variansi dari total data.

Daftar Pustaka

[1]. Hsu, Y. L., P. Y. Huang, dan D. T. Chen., 2014. Sparse Principal Component Analysis In Cancer Research. *Transl Cancer Res.* 3(3): 182–190.

[2]. Soemartini., 2012. *Aplikasi Principal Component Analysis (PCA) dalam Mengatasi Multikoliniearitas Untuk Menentukan Investasi Di Indonesia Periode. 2001.1 – 2010.4*. Bandung: Universitas Padjajaran.

[3]. Rachmatin, D., 2015. Aplikasi Metode Weighted Principal Component Analysis (WPCA) dengan Software S-PLUS2000. *Jurnal Penelitian Sains UNSRI.* 17(2): 51-58.

[4]. Sharma, S., 1996. *Applied Multivariate Techniques*. New York: John Wiley and Sons, Inc.

[5]. Varmuza, K. dan P. Filzmoser., 2009. *Introduction to Multivariate Statistical Analysis in Chemometrics*. Florida: CRC Press.

[6]. Tantular, B., 2011. Praktikum Analisis Data Multivariat II Menggunakan Software R: Modul 1 Analisis Komponen Utama. *https://berthoveens.files.wordpress.com /2011/07/modul-multi.pdf* dan *bertho@unpad.ac.id*. 20 Desember 2018(09:17).

[7]. Hair, J.F. Jr. , Anderson, R.E., Tatham, R.L., dan Black, W.C. 1998. *Multivariate Data Analysis, (5 th. Edition)*. Upper Saddle River, NJ: Prentice Hall.
[8]. Zou, H., T. Hastie, dan R. Tibshirani. 2004 *Sparse Principal Component Analysis*. California: Department of Statistics, Stanford University.
[9]. Ramadhini, Fitri. 2014. *Penyusutan Koefisien dan Seleksi Variabel Regresi dengan Elastic Net* [skripsi]. Yogyakarta: UGM.
[10]. Cadima, J. F. dan I. T. Jolliffe, I. T. 1995. Loadings and Correlations in the interpretation of *Principal Components. Journal of Applied Statistics*. 22(2): 203-214.
[11]. Zou, H., T. Hastie, dan R. Tibshirani. 2006. Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics.* 15(2): 265–286.
[12]. Zou, H. dan T. Hastie. 2003. *Regression Shrinkage And Selection Via The Elastic Net.* California: Department of Statistics, Stanford University.
[13]. Badan Pusat Statistik. 2015. *Data Indikiator Kemiskinan Penduduk Indonesia Tahun 2015* [ditemukan pada situs http://www.bps.go.id pada 13 Februari September 2018].